Europäisches Patentamt
European Patent Office
Office européen des brevets

Bescheinigung Certificate Attestation

Die angehefteten Unterlagen stimmen mit der ursprünglich eingereichten Fassung der auf dem nächsten Blatt bezeichneten europäischen Patentanmeldung überein.

The attached documents are exact copies of the European patent application described on the following page, as originally filed.

Les documents fixés à cette attestation sont conformes à la version initialement déposée de la demande de brevet européen spécifiée à la page suivante.

Patentanmeldung Nr. Patent application No. Demande de brevet n°

00125608.0

CERTIFIED COPY OF PRIORITY DOCUMENT

Der Präsident des Europäischen Patentamts;
Im Auftrag

For the President of the European Patent Office

Le Président de l'Office européen des brevets
p.o.

I.L.C. HATTEN-HECKMAN

DEN HAAG, DEN
THE HAGUE, 14/05/01
LA HAYE, LE

Europäisches Patentamt
European Patent Office
Office européen des brevets

Blatt 2 der Bescheinigung
Sheet 2 of the certificate
Page 2 de l'attestation

Anmeldung Nr.:
Application no.: 00125608.0
Demande n°:

Anmeldetag:
Date of filing: 23/11/00
Date de dépôt:

Anmelder:
Applicant(s):
Demandeur(s):
International Business Machines Corporation
Armonk, NY 10504
UNITED STATES OF AMERICA

Bezeichnung der Erfindung:
Title of the invention:
Titre de l'invention:
Method for improving data retrieval in large collections of data

In Anspruch genommene Priorität(en) / Priority(ies) claimed / Priorité(s) revendiquée(s)

Staat: State: Pays:	Tag: Date: Date:	Aktenzeichen: File no. Numéro de dépôt:

Internationale Patentklassifikation:
International Patent classification:
Classification internationale des brevets:
/

Am Anmeldetag benannte Vertragstaaten:
Contracting states designated at date of filing: AT/BE/CH/CY/DE/DK/ES/FI/FR/GB/GR/IE/IT/LI/LU/MC/NL/PT/SE/TR
Etats contractants désignés lors du depôt:

Bemerkungen:
Remarks:
Remarques:

D E S C R I P T I O N 23 Nov. 2000

Method for Improving Data Retrieval in Large Collections of Data

1. Background of the Invention

1.1 Field of the Invention

The present invention relates to a method, corresponding means and a corresponding program product for improving data retrieval in large collections of documents. Specific focus is given to improvements of the search for information within Web pages making up the Internet.

1.2 Description and Disadvantages of Prior Art

Today's world is characterized by that of a "connected community"; for the business world it is called "e-business", for everyday people simply "the Internet". One of the most important functions/services used are **search engines**. These provide services aimed at finding information requested by users or applications.

Until a few months ago, search engine providers were pushing their search/data retrieval technology to allow to index "the whole Internet". These approaches struggle with several limitations. Due to the tremendous growth rate of the number and sizes of Web pages it has become very problematic for these technologies to provide the required processing power and the required storage to create and maintain the search indexes. Moreover, a typical search pattern will result in too many search hits and the result itself cannot be analyzed anymore, that is, even the complexity of the search results became a problem. The reason for this is that most of the retrieved documents (though containing the search pattern) do not have any semantic relationship to the intended notion behind the search pattern; that is, most of the retrieved documents are just irrelevant.

As of today, search technologies are competing to find "documents" relevant to the information need of a user, thus focusing on quality rather than quantity. This is an important factor which allow for information services within corporations such as knowledge management services.

Search engine technology today is a straight forward process, involving technologies that had been well known for years. This technology can be described as follows:
Documents, in this case Web pages of the Internet, are collected by a sort of Web crawler and are processed so that their contents can be stored in a fulltext **index** (typically an **inverted index**). The process is basically to build a list of keywords plus their references to documents they were found in (inversion); that is, keywords plus positional information allowed to locate an indexed keyword or token within the processed documents. This requires to split the document into informational "units", which are composed of single words, and to record their positional information (that is, the documents they appear in and their position(s) within the documents) within the index. The "keys" that are stored, are the words pointing to documents together with the associated positional information (this is the inversion process). Finally, the information available in the index is exploited by later search queries to match search queries against the collection of indexed documents. The search result is a list of documents representing possible document candidates relating to the search pattern.
As the result list comprises so many document candidates with almost no semantic relationship to the "Concept" or "Notion" behind the search pattern, further technologies have been applied providing additional information to a user. For instance the resulting documents can be scored in an order which represents their "relevance" to the query, taking into account the occurrences of words in the collection, and the occurrences

of words in a document. These technologies are available for the example under the terminology of "relevance ranking" and "probabilistic ranking". Other approaches apply for instance "popularity scores" for documents, based on how frequently they are referenced or had been visited/selected by other users. These popularity scores are then used for the ranking process of the list of result documents.

Whichever combination of above mentioned technologies are selected, severe disadvantages adhere to any approach.
The relevance (independent of the type of relevance measure) of the retrieved documents is mostly regarded as being poor. Therefore, users typically need to issue more than one search request to find (if at all) the information they are seeking. This iterative approach (of narrowing a search and thus the result set of documents) slows down the search process significantly. In any case the really relevant documents within a search result list are embedded ("buried") in an often very large number of other, non-relevant documents (as judged by the user in an ex-post analysis).

Above mentioned problems still further increase as the number of documents accessible via the Internet and the storage requirements to cope with this flood of data will increase dramatically in the near future. The following observations will introduce to this challenge.
Information technology is moving towards a world, that has opened its knowledge base via the Internet. In addition, companies a moving to be connected as well (the e-business paradigm). One main struggle is providing technology that allows to manage information and knowledge for the purpose of making it available to users on time and with optimal quality. Until a few years ago, the amount of information being made available was influenced by cost of storage (hard disks) and the hardware required to allow the operation of such information systems. With the storage cost decreasing, tendencies to store

information electronically (including archiving) has increased dramatically. This all puts a burden on data/information management systems and exploiting applications.

Reflecting this to the search engine world, search engine technology needs to manage huge amounts of data. The technology is further burdened by the requirement to seek through the thus stored information in a fraction of a second and to return relevant information.

Thus the severe problems the technology has to cope with relate to the challenge of improving the quality of search as well as the management of terabytes of data required to support the search technology.

A short scenario which could be based on any of the well-known Internet search engines such as "Fast!", "Thunderstone", "Altavista", "Google", explains the deficiencies of the current search technology in more detail. Anyone who has used these, knows that it can be painful and time consuming to find the "right information"; typically one shot at the search engine is not enough.

The example scenario, is a person trying to find information about the inkjet printer he has at home, which might be related to the device driver he has currently installed. Suppose, the user does not have any specifics other than that when he tries to print a page, the printer lights flash, then nothing happens. He might then consult his manufacturer's web site, but doesn't find an answer; the driver he has is the latest version. The next step is then to check on the Internet. He then enters the following query in the search field of a search engine (like Altavista):

"problem with Epson color inkjet"

What does the search engine do with this information? First of all the search engine does not know that this is a problem

statement. Essentially, it takes/isolates the query string into single words (optionally it could drop "trivial words" such as "with") and then locates those documents were each of these occur. Given the immense size of the Internet and the capacity that, e.g. Altavista can handle (200 million documents), it is obvious, that each of the words occur in a huge amount of documents (> 200.000). Assuming that the common set of documents is still in the range of 10.000 documents, it is obvious, that a user cannot browse through all of these. Thus the next step is for the search engine to figure out which search hits are the most relevant ones ! This is the problem! According to the state of the art relevance can be determined using underlying algorithms that take into account the information available in the fulltext index and the search terms used. The processing for this example scenario is straight forward:

1. for each candidate document the occurrences of each search term is determined;
2. given this information a rank score for each document (e.g. the normalized sum of the occurrences) is calculated;
3. once the candidate list of documents has been completely processed, the document list is sorted by descending rank scores; and
4. the ranked list of documents is returned to the user.

Though the retrieved documents of above example search scenario do contain the words specified in the search request, a further analysis of the results leads to the following observations:

a. The words comprised by the search pattern do not occur in the requested/intended context. The retrieved documents almost never actually do mention "problems with Epson printers".

b. If the retrieved documents even relate to problems with Epson printers these documents comprise sentences, which are variants of the following: "A problem with the Epson XYZ color printer is not known."! This search hit is actually a algorithmically determined "close" match with the query but from a semantic point of view actually addresses a completely

different context !

c. Depending on the information presented in the search request (the search pattern), very often the used vocabulary consists of commonly used/found words. The result is that in the list of the retrieved documents it is almost impossible to distinguish with respect to the importance of the retrieved documents. Search pattern with search terms which are not very "selective" typically result in a list of retrieved documents with rank scores which are very similar, that is which do not show a strong variation ("density of rank scores"). The consequence is: In such a circumstances rank values are inappropriate means to distinguish between important and unimportant documents.

d. Finally these problems increase at the same pace as the data volume increases.

1.3 Objective of the Invention

Thus the invention is based on the objective to provide a technology which improves the "Quality of search" in terms of retrieving documents which are more relevant in view of the semantic "concept or notion" represented by the search pattern.

It is a further objective of the current invention to reduce the storage requirements of information structures supporting the data retrieval technology in identifying and locating documents representing potential search hits in view of a search pattern.

Finally, the current invention has the objective to improve the required processing time for processing individual data retrieval requests.

2. Summary and Advantages of the Invention

The objectives of the invention are solved by the independent claims. Further advantageous arrangements and embodiments of the invention are set forth in the respective subclaims.

The present invention relates to a method, corresponding means and a corresponding program product for improving data retrieval in large collections of documents.
The invention suggests a technology for automatically creating a search index usable by a later query on a multitude of documents. In a first functional component or step a document to be indexed in said search index is retrieved. In a second functional component or step a document extract from said document comprising a portion of said document only is generated. Finally in a third functional component or step said search index is enhanced based on said documents extract as substitute of said document.

The suggested approach significantly improves the "Quality of search" in terms of retrieving documents which are more relevant in view of the semantic "concept or notion" represented by the search pattern. Moreover the storage requirements of information structures supporting the data retrieval technology in identifying and locating documents representing potential search hits in view of a search pattern are reduced and at the same time the required processing time for processing individual data search requests can also be reduced.

3. Brief Description of the Drawings

Figure 1 provides an overview of the current state of the art search technology by visualizing the basic system structure and the data flows in connection with an search request.

Figure 2 illustrates in contrast to Fig. 1 the "Extractor" technology according to the current invention imbedded into the state of the art environment and visualizes its influence on system resource requirements (like hard disk space and data throughput for the index build process).

4. Description of the Preferred Embodiment

In the drawings and specification there has been set forth a

preferred embodiment of the invention and, although specific terms are used, the description thus given uses terminology in a generic and descriptive sense only and not for purposes of limitation. It will, however, be evident that various modifications and changes may be made thereto without departing from the broader spirit and scope of the invention as set forth in the appended claims.

The present invention can be realized in hardware, software, or a combination of hardware and software. Any kind of computer system - or other apparatus adapted for carrying out the methods described herein - is suited. A typical combination of hardware and software could be a general purpose computer system with a computer program that, when being loaded and executed, controls the computer system such that it carries out the methods described herein. The present invention can also be embedded in a computer program product, which comprises all the features enabling the implementation of the methods described herein, and which - when being loaded in a computer system - is able to carry out these methods.

Computer program means or computer program in the present context mean any expression, in any language, code or notation, of a set of instructions intended to cause a system having an information processing capability to perform a particular function either directly or after either or both of the following a) conversion to another language, code or notation; b) reproduction in a different material form.

Even though the current invention is described within the context of the search problem in the "Internet" this is for a descriptive purposes only and may not be understood as a limitation of the applicability and scope of protection of the current technology. The search problem in the Internet has been selected as the representative for the search problem within even greater repositories of information in the form of

"electronic documents" found in many of the large enterprises/organizations. These repositories may easily surpass the current size of the Internet (2 to 8 Terabytes of data) in terms of the comprised number of documents and the amount of occupied storage.

To a certain extent the current invention is illustrated based on Data/Text Mining technology offered by IBM's "Intelligent Miner" product family. Also with respect to these aspects no limitation of the current invention may be deduced as other mining technologies may be exploited instead.

4.1 Introduction

Fig. 1 provides an overview of the current state of the art search technology by visualizing the basic system structure and the data flows in connection with a search request.
To offer a search service 100 on some display device 101 first a complex infrastructure has to be set up within which the search request will be processed. Web crawlers 103 retrieve documents from the network 104 and store the retrieved documents within some temporary document store 105. In a further, separate step a further component, called the **indexer** 106, parses the documents within the temporary document store into individual keywords, associates the keywords with positional information referring to their locations within the individual documents and enhances the **search index** 107 (an inverted index) with this information. When a search request, rudimentary depicted by arrow 108, is entered into the system, the search pattern is used to search the index only (on behalf of the large collection of documents) and the (potentially ranked) list of search hits is returned.

4.2 The Information Extractor and a New Type of Search Index

The fundamental observation according to the current invention to cope with the current problem of improving search technology as to manage huge amounts of information effectively, both in terms of quantity and quality, is, that all these difficulties

are caused (at least to a large extent) by the contents of the search index according to the state of the art. All the problematic information, which will during an actual data retrieval process misguided the search engine, has been incorporated into the search architecture already when the search index has been created. It is thus the observation of the current invention that the process of creating a search index has to be modified such that only the **characteristic portions** of a document will find their way into the search index.
Put in a nutshell the current invention therefore suggests a new component called "**Information Extractor**", which uses a document as input and generates a so-called **document extract** which comprises a portion of said document only. The document extract is generated such that it is most characteristic for the document as a whole. For those portions comprised by the document extract also positional information will be included referring to positions within the original document. As will be outlined below it is suggested to use data mining technology for this purpose of generating a document extract. Finally the search index will be created/enhanced not based on the document itself but on the document extract.
This results in the achievement that no information misguiding the search engine will enter the search index anymore. Thus only those portions of a document which are really characteristic to this document will participate within the search process and allow a search pattern to "Hit" this document. Moreover this teaching allows to significantly reduce the size of the search index, which in addition will speed up the search process significantly.

Fig. 2 illustrates in contrast to Fig. 1 the proposed "Extractor" technology according to the current invention imbedded into the state of the art environment of Fig. 1 and visualizes its influence on system resource requirements (like hard disk space and data throughput for the index build process). The process of building the new type of fulltext index

supporting a search engine in the later query involves the following functional components or method steps:

1. Within step 203 the functional ability to gather documents from various types of document repositories ("pull" technology) or alternatively allow services to send a notification which document(s) (their identification) should be processed, or even send the complete documents themselves ("push" technology) has to be available. Typically these functionalities is are provided by so-called Web crawlers.

2. Within step 209 a new document analysis component, called here an "**information extractor**" or more picturesque a "**condenser**" is suggested. As explained above the output of the extractor is a new virtual document, the document extract, whose contents describes the condensed information extracted by the extractor's analysis procedure from the original document. The document extract also comprises positional information referring for the contents of document extract to its occurrence within the original document. Typically the document extract comprises a portion of the original document only generated for instance by data mining technology by a selection process from the original document. On the other hand it is also possible to apply "document understanding" technology to determine the document's semantic and to generate an "abstract" of the analyzed document, which would go beyond a mere selection process for text elements. The document extract is intended to be the replacement for the original document, which is then processed further. The "condensed" document extract is then stored in a temporary document store 205 on behalf of the original document. The original document is not required anymore and can be discarded.

As indicated by Fig. 2 in comparison to Fig. 1 the amount of storage requirements (symbolized by the number of hard disk drive icons) for the temporary document store 205 vs 105 can significantly be reduced as it is required to store the much smaller document extract only. This increased storage efficiency

pays back by allowing to store a greater amount of information, before reaching capacity limits (refer also to the table given below).

3. In the next step the indexer 206 decomposes the (virtual) document extract into a set of "words/keywords/tokens" that are then stored together with their positional information in a fulltext index 207, which forms the basis for the actual search engine. As indicated by Fig. 2 in comparison to Fig. 1 the amount of storage requirements for the search index 206 vs 106 can significantly be reduced as it is required to store the index information of the much smaller document extract only.

4. Finally the search service itself (symbolized by an arrow 208) allows to issue queries against the fulltext index and returns the (optionally ranked) result list back to the requesting client/user 200.

4.3 The Technology Exploited by the Preferred Embodiment of the Extractor

The extractor according to the current invention analyzes a document for its informational content suppressing all those portions of a document deviating from its actual topic or theme; thus the extractor could be viewed as an instrument for the determination of a document's "relevance". In the state of the art approach, the notion of relevance of a document enters the search process at a very late stage, namely during the ranking process. Moreover the notion of relevance is determined in conjunction with a search pattern only, whereas the current invention uses a relevance approach with a scope limited to the document only. This explains why in contrast to the state of the art different technologies are exploited for the relevance determination. The proposed extraction process takes place already when creating the information structures (especially the search index) supporting the data retrieval process. This is contrary to the state of today's technology, where the relevancy is determined at runtime of the search request (at this point in time the index is already built and thus static).

Summarization Technology from the Area of Text Mining

The relevant information to be incorporated into the document extract can be determined by those sentences or parts of sentences in the document that actually contain the relevant and descriptive keywords. The area of data mining provides technologies for automatically generating from a certain document a so-called **summary** or **abstract** comprising the most relevant (according to the document's semantic) portion of said document only. IBM's Intelligent Miner product family offers such technologies as one example. The current invention suggests to exploit this technology to generate a document extract.

A document summary, used as document extract according to the current invention, consists of a collection of sentences extracted from the document that are characteristic of the document content. A summary can be produced for any document but it works best with well-edited structured documents. Based on certain control parameters one even can specify the maximum number of sentences the summary should contain, either as an absolute number or in proportion to the length of the document. Typical summarization tools use a set of ranking strategies on word level and on sentence level to calculate the relevance of a sentence to the document. The sentences with the highest scores are extracted to form the document summary. One even can tune the way the relevance is calculated by varying threshold parameters
and coefficients in certain configuration files.

To just give an example the following document:

> BANGALORE, India, M2 PRESSWIRE via Individual Inc. : AT&T today launched India's first Global Network Management Centre (GNMC) to meet the networking needs of local companies and multinational corporations (MNCs) in India. AT&T will provide advanced network solutions, as well as a

range of sophisticated communications services, to large Indian companies and domestic and foreign MNCs country-wide. The GNMC will be located in Bangalore. The state-of-the-art facility is connected to AT&T's other GNMCs in China, Singapore, the United States and Europe. The facility uses the latest communications technology to manage, maintain and operate customers' networks 24-hours-a-day, 365 days-a-year. "The Bangalore GNMC shows our commitment to providing local and global customers with world-wide network management capabilities," said Joydeep Bose, director, AT&T Managed Network Solutions, India. "This facility is a significant technological investment and is the first-ever of its kind in the country."

The GNMC will be run by AT&T's Managed Network Solutions division, which focuses on the communications needs of MNCs world-wide. AT&T will also offer an extensive, flexible range of communications services including network analysis and design, network integration and implementation, and a complete suite of outsourced network operations management services. AT&T Managed Network Solutions will provide world-class, product-independent services for voice and data networking to help customers choose the best technology and transmission facilities the market can offer.

"More and more companies are setting up or expanding their businesses in India," said Rakesh Bhasin, president, AT&T Managed Network Solutions, Asia/Pacific. "In order to expand efficiently, they need communications networks they can trust. AT&T can help save companies time, money and resources by offering expert advice on installing and 'future proofing' a network, managing it once it has been built, and making sure it provides consistent, high-quality, seamless voice and data connections."

will be summarized by the summarization technology provided by IBM's Intelligent Miner product family into:

BANGALORE, India, M2 PRESSWIRE via Individual Inc.: AT&T today launched India's first Global Network Management Centre (GNMC) to meet the networking needs of local companies and multinational corporations (MNCs) in India. The GNMC will be run by AT&T's Managed Network Solutions division, which focuses on the communications needs of MNCs world-wide.

Extracting Characteristic Sentences, Parts of Sentences, (Key)Words or Tokens in General Using Mining Technology

The current teaching suggests to exploit various other technologies from the area of data mining alone or in combination with one another.

To generate the document extract comprising a portion of the document only certain (key)words occurring within the document may be extracted based on word ranking approaches.
The complete document is analyzed, however, not all words in a document are scored. Typically words must fulfill one of the following criteria to be eligible for scoring:

a. The word appears in certain document structures, such as titles, headings, or captions.

b. The word occurs more often in the document than in the document collection represented by a reference vocabulary. This is known as the **word salience measure**.

c. The word must occur more than once in the document.

The generated score of a word consists of the salience measure if this is greater than a threshold set in the configuration file. The default **salience measure** is **(text frequency) * (inverse document frequency)**, or **(tf*idf)**. Moreover further weighting factors may be introduced if a word occurs in the title, a heading, or a caption or other specific syntactical locations within a document.

To generate the document extract comprising a portion of the

document only certain sentences or parts of sentences occurring within the document may be extracted based on sentence ranking approaches.
Sentences in a document are scored according to their relevance to the document and their position in a document. The sentence score may be defined as the sum of:

a. The scores of the individual words in the sentence multiplied by a coefficient set in the configuration file.

b. The proximity of the sentence to the beginning of its paragraph multiplied by a coefficient set in the configuration file.

c. Final sentences in long paragraphs and final paragraphs in long documents receive an extra score.

d. The proximity of a paragraph to the beginning of the document multiplied by a coefficient set in the configuration file.

The highest ranking sentences are extracted to create the document summary. One also can specify the length of the summary to be a number of sentences or a percentage of the document's length.

Alternatively, a keyword list (e.g. domain specific words) can be used to extract those parts/words of the document that are in close proximity to each of the listed keywords, thus focusing on a subset of documents.

Feature Extraction

To generate the document extract comprising a portion of the document only so-called "**Features**" occurring within the document may be extracted based on **feature extraction** technology.
Many of the technologies and tools developed in information mining are dedicated to the task of discovery and extraction of information or knowledge from text documents, called feature extraction. The basic pieces of information in text--such as terms made up by a collection of individual the words like for instance company names or dates mentioned--are called features.

Information extraction from unconstrained text is the extraction of the linguistic items that provide representative or otherwise relevant information about the document content. These features a can be extracted or used to assign documents to categories in a given scheme, group documents by subject, focus on specific parts of information within documents. The extracted features can also serve as meta data about the analyzed documents.

The feature extraction component of IBM's Intelligent Miner product family recognizes significant vocabulary items in text. The process is fully automatic -- the vocabulary is not predefined. When analyzing single documents, the feature extractor can operate in two possible modes. In the first, it analyzes that document alone. In a second preferred mode, it locates vocabulary in the document which occurs in a dictionary which it has previously built automatically from a collection of similar documents. When using a collection of documents, the feature extractor is able to aggregate the evidence from many documents to find the optimal vocabulary. For example, it can often detect the fact that several different items are really variants of the same feature, in which case it picks one as the canonical form. In addition, it can then assign a statistical significance measure to each vocabulary item. The **significance measure**, called "**Information Quotient**" (IQ), is a number which is assigned to every vocabulary item/feature found in the collection; e.g. features that occur more frequently within a single document than within the whole document collection are rated high. The calculation of IQ uses a combination of statistical measures which together measure the significance of a word, phrase or name within the documents in the collection.

Based on above mentioned technologies the extractor can even determine whether or not there is relevant information within the document at all (e.g. using threshold values); that is, the extractor is even able to determine documents without any relevance for which then no document extract would be generated.

This can be achieved by using/providing threshold information regarding the quality of the generated summary or informational content. The generated document extracts as a whole can have a relevance score assigned (based on its considering sub-components), denoting how well they describe the contents of a document. In a range of 1 .. 100, one could assume, that values above a certain threshold (for instance 75%) are good descriptors of the overall document. This knowledge can be used to determine whether or not a document should be stored in the fulltext index at all; only if the threshold will do is reach a document extract will be used to enhance the search index. For instance a document identified as "John Doe's home page" is most likely of no interest to the global Internet community. So it is a candidate to drop completely.
A similar problem relates to "spaming" in the Internet, which refers to introducing a huge amount of data actually not related to the web site at all just to increase the probability of being found by many "typical search requests". The current invention would automatically detect such documents with the result that these documents are not being considered to be stored in the fulltext index.

4.4 Integrating the Extractor Within a Data Retrieval Architecture

With respect to incorporation of the extractor within an existing data retrieval architecture several possibilities will be suggested. It is important to note, that the incorporation of the extractor within the existing system providing search capabilities, can be done without (or only little) changes to the existing architecture.

The new extractor component hooks into an existing search system at the point between the physical fetching of a document from a document repository (e.g. the Internet or a document management system like an electronic library) and the point where the token

list for a document is inserted into the fulltext index (basic search indexer functionality).
With respect to Fig. 2 the following three options for enabling the information extractor component are suggested:

1. the extractor can be incorporated as an extension of the process of fetching a document 203 (e.g. a web crawler (pull technology) or a push agent). This approach is visualized within Fig. 2 as implementation option 1. It has the significant advantage of reducing the storage requirements for the temporary document store 205 as only the much smaller document extract instead of the original document has to be stored.

2. the extractor can be incorporated as a daemon process which manipulates the documents that are temporarily stored on disk 205 before the search index is enhanced by the indexer 206. For that purpose the extractor could be invoked for instance by file system notification services. This approach is visualized within Fig. 2 as implementation option 2.

3. the extractor can be incorporated as an additional document analysis based process invoked as a preprocessing phase to the indexer, before tokenization of document(s) is performed. For this purpose the extractor could be invoked by the indexer 206. This approach is visualized within Fig. 2 as implementation option 3.

4.5 A Reference Implementation

An implementation of above mentioned teaching has been developed using standard IBM software; specifically the IBM Intelligent Miner for Text has been exploited.

As the collection of documents not the web pages in the Internet but IBM's internal IBM intranet and its documents have been chosen. The robot to retrieve the documents was the IBM web crawler. The documents were fetched and stored temporarily on hard disks. The "information extractor" was implemented as a "stand-alone" application, which was directed to a subset of the temporarily stored documents, for which in return it created a

three sentence summary. The documents that had been thus condensed, were replaced with the summary as document extract. After completion of this task, the indexer picked up the condensed version of the documents and indexed these.

Of course extra processing time has to be spent for the execution of the extractor, that is for the "summarizer process". But compared to the state of the art approach wherein the whole original document would be tokenized to enhance the search index accordingly this extra processing overhead for the extractor is overcompensated by the fact that to process the document extract far less data is to be analyzed by the indexer overall. Therefore, already the indexing performance has been improved significantly. Due to the much smaller word list (since the vocabulary is much more controlled) or in other words of the much smaller search index and their associated document reference list the overall search performance could be improved even more.

Based on the experience of this reference implementation an extrapolation to the whole Internet results in an efficiency calculation reflected by the following table:

Traditional		
Total number of documents	900 000 000	documents
average size of a document	5120	bytes
traditional search engine	4291,53	GBytes data to be processed
resulting index size	1502,04	GBytes index size
New		
relevant documents in total (20%)	180 000 000	documents
output of the "information condenser"	512	bytes
information condenser" enabled search engine	85,83	GBytes data to be processed
resulting index size	30,04	GBytes index size

As can be seen from this table extractor based search service according to the current invention requires only 30 GBytes disk space for its fulltext index, opposed to the traditional search engines 1500 GBytes ! These numbers make obvious that based on the state of the art technology the a single high end server cannot handle this amount of data. Therefore an Internet search service today is based on a cluster of typically more than 50 of these servers. On the other hand, with an index of merely 30 GBytes in size generated according on the current invention it appears to be possible to host such a search services on a single high-end server!

Based on the experience with the reference implementation according to the current invention it is interesting to note that only about 20% of the analyzed documents seem to be relevant at all resulting in the generation of a document extract.

4.6 Advantages of the Invention

Besides improvements of the current invention with respect to storage requirements and processing time for individual search requests the current invention improves the quality of the search results significantly; that is the relevance or precision of the returned search hits match the semantic "notion or concept" expressed by the search pattern much more accurate than traditional technology.

The advantages of current invention can be understood best by a comparison with search engines according to the state of the art returning relevant documents (relevancy displayed by rank order of the result list and optionally rank scores per document) for which optimally the first document in the list is the best match for the query, if one takes the following statistics into account: Statistics taken from big search engine installations show that 40% of the "words" indexed will never be searched for; this portion comprises artificial words, explicit numeric values (not approximations like 1999, or 1.000.000) and the like. Another 40% of words in a document are "filler words" required to "ornament" the text and the overall appearance of the text not introducing further semantics into a document. The remaining 20% can be considered to be relevant to the informational content of the document; the current invention tries to locate specially this portion of a document and will extract this into the document extract.

The search quality according to the current invention can be measured for instance by the quotient between "**recall**" and "**precision**" defined as "**relevance**" **relevance=precision/recall**. Recall specifies the number of documents returned for a given query. Precision specifies, the number of the recalled documents which are relevant to the query (in an ex post investigation). Ideally, the quotient would be 1.0, in the real world, an optimum is in the range of 0.3 to 0.5, but seldom surpassed.

The influence of "the information condensing" according to the current invention on these factors for improving the search quality can easily be understood.
The "recall" measure is decreased, by dropping documents from the index completely determined as irrelevant due to lack of information overall. Thus "coincidentally" containing a certain keyword will not occur. Therefore in general the number of documents that contain a keyword are decreased by the extractor preprocessing step beforehand, taking into account the important information a given document has to offer.
The "precision" is increased on the other hand by condensing of information for a given document by selecting the most characteristic portions of a document. Multi-word search requests will thus even more distinguish "good" from "lesser good" matches due to their close proximity (e.g. occurring in same sentence). The number of occurrences overall in the document will also indicate a higher relevancy.
In essence, as the recall decreases and the precision increases the overall quotient will grow towards 1.0 and thus improve.

The responsiveness of the search service according to the current invention will definitely benefit from the lesser amount of information needed to be looked up in the fulltext index, which is also a quality aspect of the search service.

C L A I M S

1. Computerized method for automatically creating a search-index usable by a later query on a multitude of documents,

said method comprising a first step of retrieving a document to be indexed in said search-index, and

a second step of generating a document-extract from said document comprising a portion of said document only, and

a third step of enhancing said search-index based on said documents-extract as substitute of said document.

2. The method according to claim 1,

wherein said second step generating into said document-extract positional information of said portion with respect to said document.

3. The method according to claim 2,

wherein said second step generating said documents-extract

by a summarization technology creating a summary of said document; or

by extracting tokens of said document, said tokens being characteristic to said document.

4. The method according to claim 3,

wherein said tokens are anyone or a combination of the following:

sentences or parts of sentences;

words of said document;

features extracted from said document based on feature extraction technology.

5. The method according to claim 4,

wherein in said method said characteristic of said tokens is measured by one or a combination of the following:

according to the frequency of occurrence of a word;

according to a word-salient-measure of a word;

according to the closeness of a word to the beginning of a of paragraph in said document;

according to the closeness of a word to the beginning of said document;

according to the closeness to or position within a document syntax element like a heading or a caption.

6. The method according to anyone of above claims,

wherein said document is a web-page in the Internet.

7. The method according to claim 7,

wherein said second step is executed directly after retrieving said document and before storing said document on temporary storage and storing not said document but said document-extract only on said temporary storage being used for said third step.

8. A search-index usable by a query on a multitude of documents,

said search-index create by a method according to anyone of the preceding claims 1 to 7.

9. A computer system comprising means adapted for carrying out the steps of the method according to anyone of the preceding claims 1 to 7.

10. A data processing program for execution in a data processing system comprising software code portions for performing a method according to anyone of the preceding claims 1 to 7 when said program is run on said computer.

11. A computer program product stored on a computer usable medium, comprising computer readable program means for causing a computer to perform a method according to anyone of the preceding claims 1 to 7 when said program is run on said computer.

A B S T R A C T

The present invention relates to a method, corresponding means and a corresponding program product for improving data retrieval in large collections of documents.

The invention suggests a technology for automatically creating a search index usable by a later query on a multitude of documents. In a first functional component or step a document to be indexed in said search index is retrieved. In a second functional component or step a document extract from said document comprising a portion of said document only is generated. Finally in a third functional component or step said search index is enhanced based on said documents extract as substitute of said document. (Fig. 2)

(Drawings)

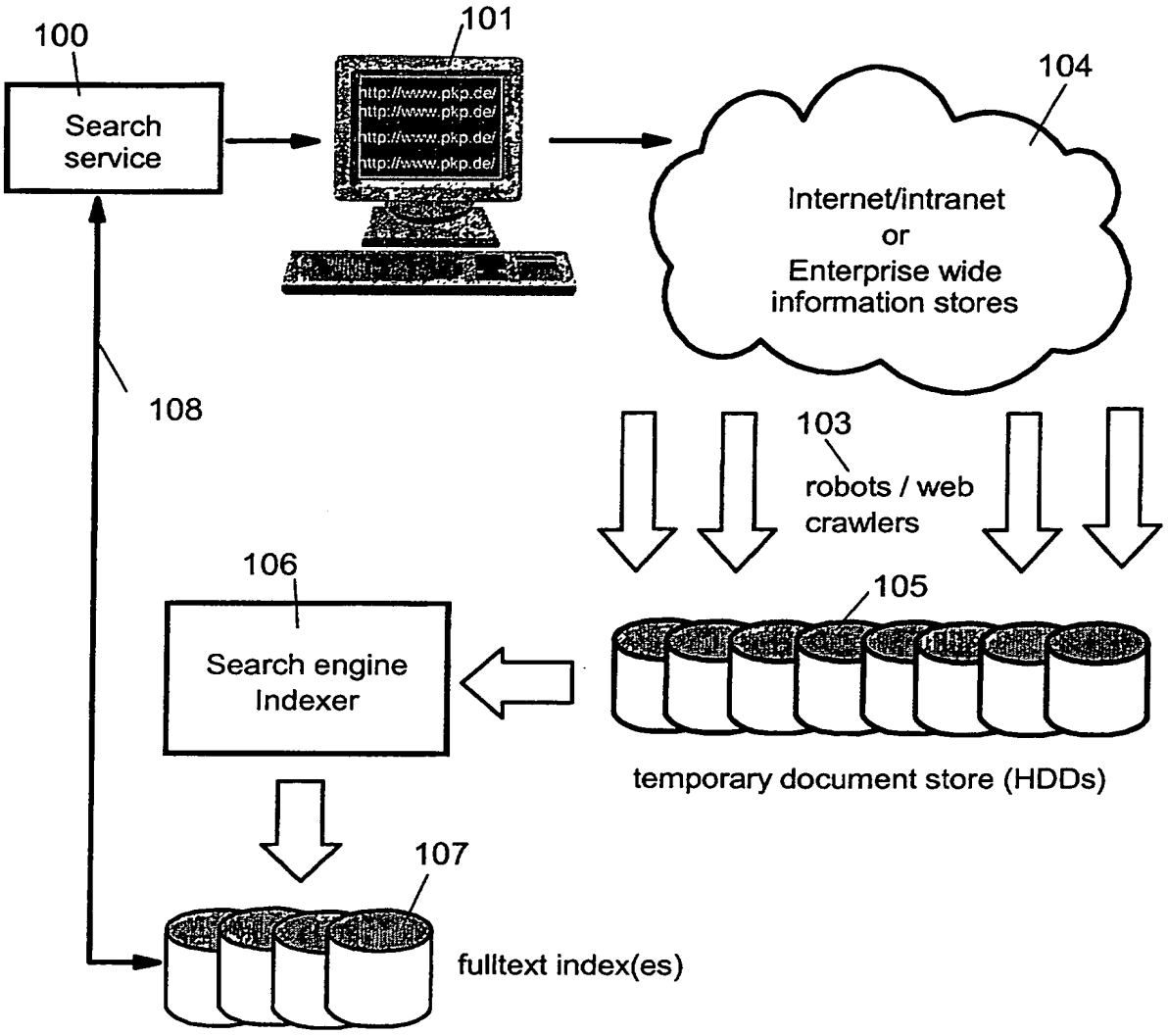

FIG. 1

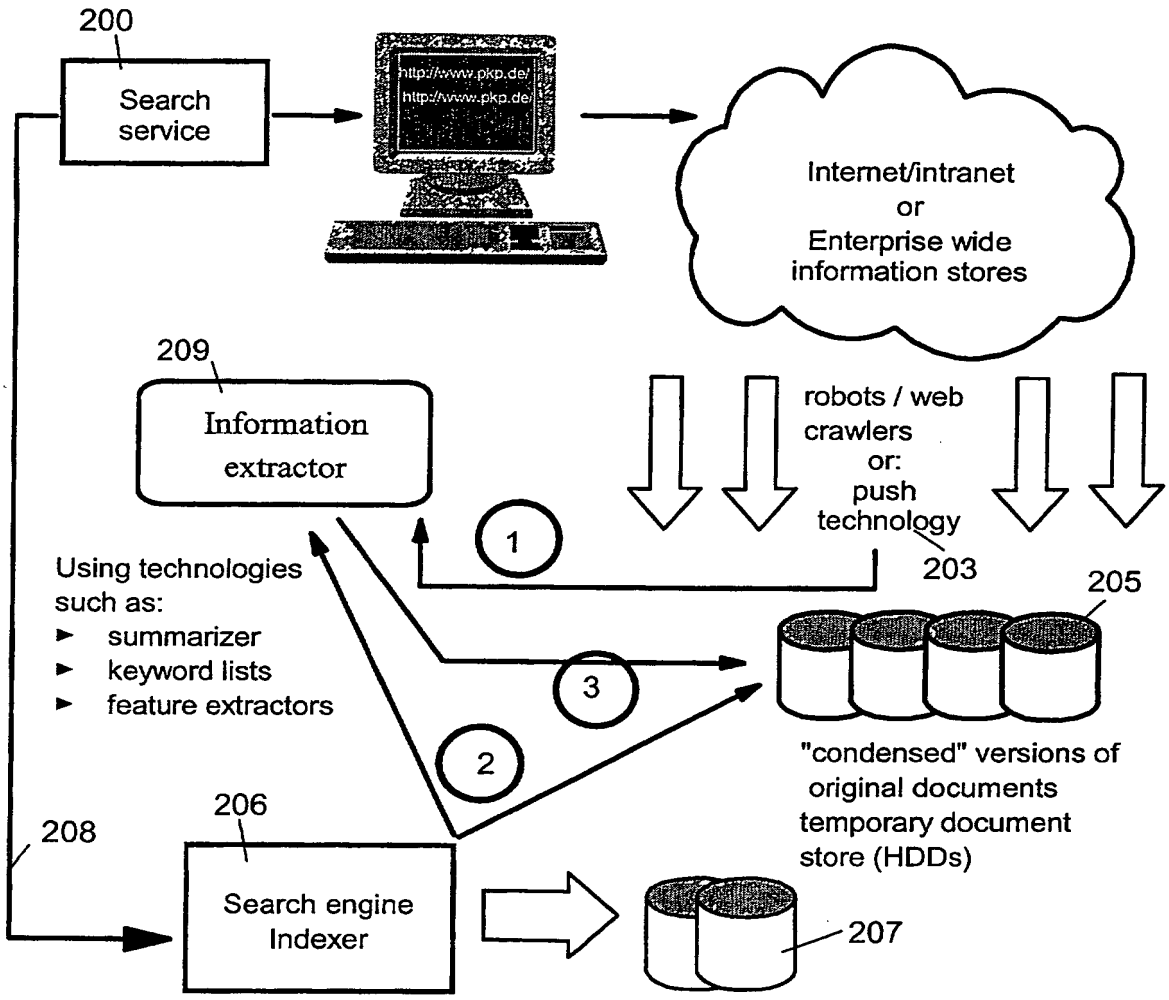
200
Search service
http://www.pkp.de/
http://www.pkp.de/
Internet/intranet or Enterprise wide information stores
209
Information extractor
robots / web crawlers or: push technology
203
205
1
Using technologies such as:
summarizer
keyword lists
feature extractors
3
2
"condensed" versions of original documents temporary document store (HDDs)
206
208
Search engine Indexer
207

FIG. 2